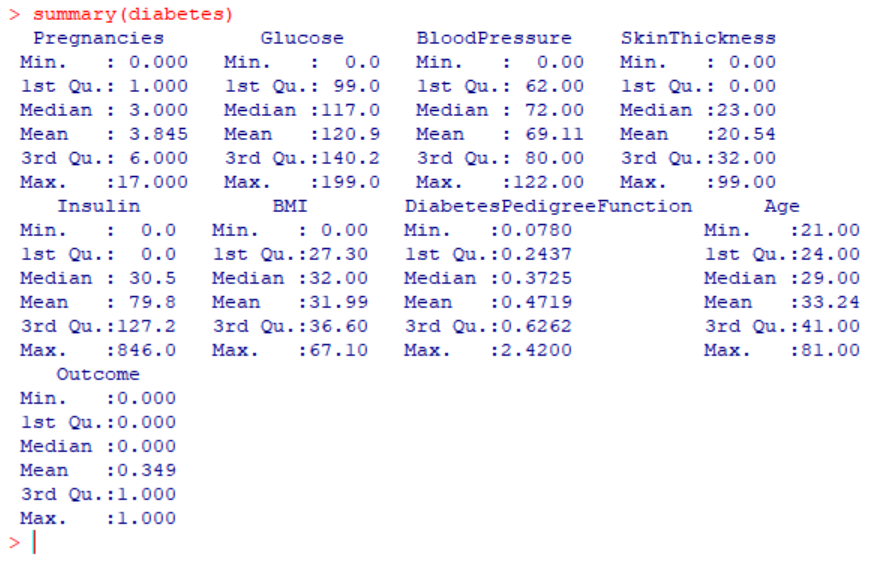
당뇨병 원인 분석

이은주, 나유정

분석 목적 : 당뇨병의 원인을 찾고 발병을 예측하기

데이터 정리 및 변수 설명



pregnancies : 임신 횟수

glucose : 경구 포도당 내성 검사에서 혈장 포도당 농도

blood pressure : 혈액 압력(확장기 혈압)

skin thickness : 삼두근 피부 두께

insulin : 혈청 인슐린

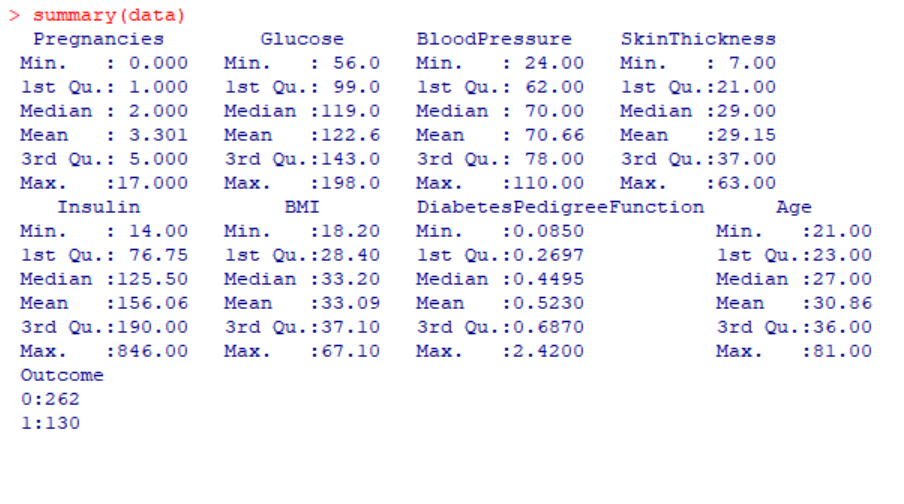
BMI : 체질량 지수(체중/(신장)^2)

diabetes pedigree function : 당뇨병 혈통 기능

age : 나이

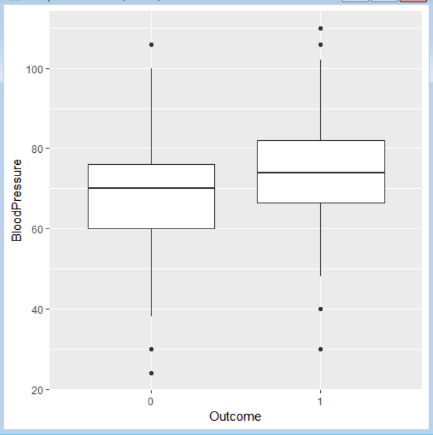
outcome : 결과(0 또는 1)

#불필요한 자료 제거



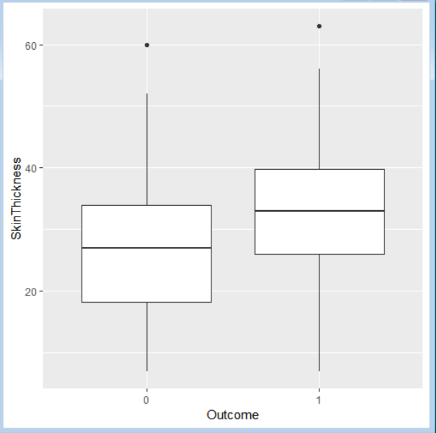
#변수와 당뇨병과의 관련성 비교

#혈압



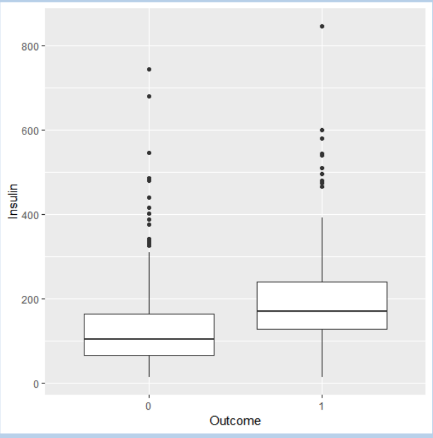
중앙값이 거의 비슷하여 큰 차이가 없어 보입니다.

#피부 두께



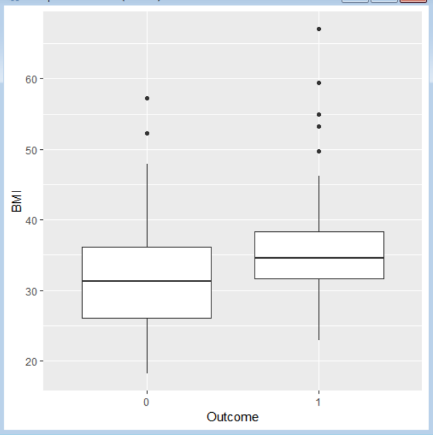
박스플랏 모양이 거의 비슷하기에 큰 차이가 없다고 생각할 수 있습니다.

#인슐린



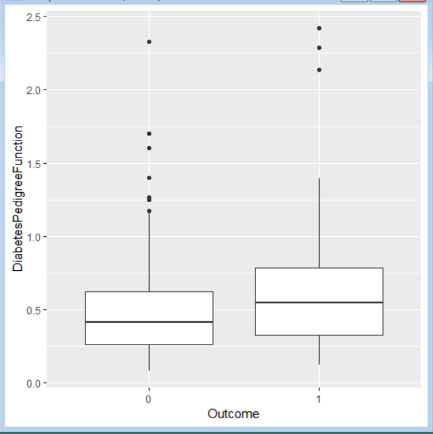
인슐린에 따른 당뇨병 발생은 큰 차이가 없어보입니다.

#BMI



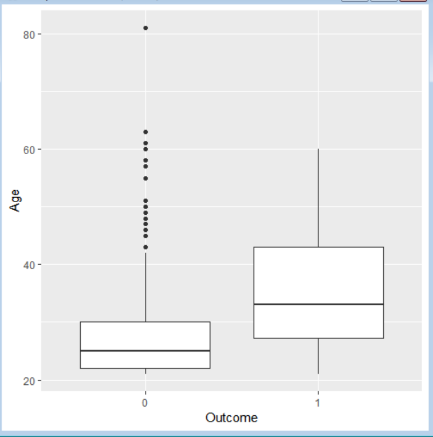
BMI에 따른 당뇨병 발생은 큰 차이가 없어보입니다.

#당뇨병 혈통 기능



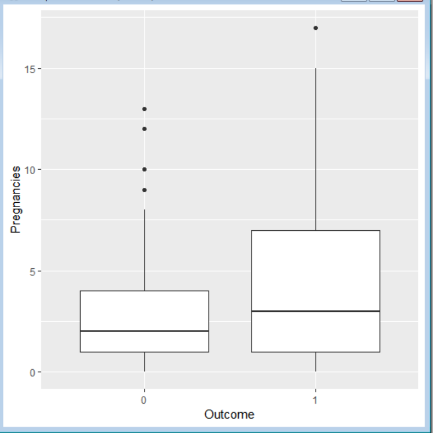
당뇨병은 혈통에 큰 영향을 받지 않습니다.

#나이



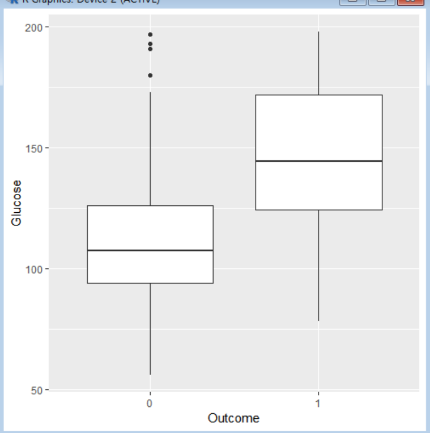
나이가 많을수록 당뇨병일 확률은 높습니다.

#임신횟수



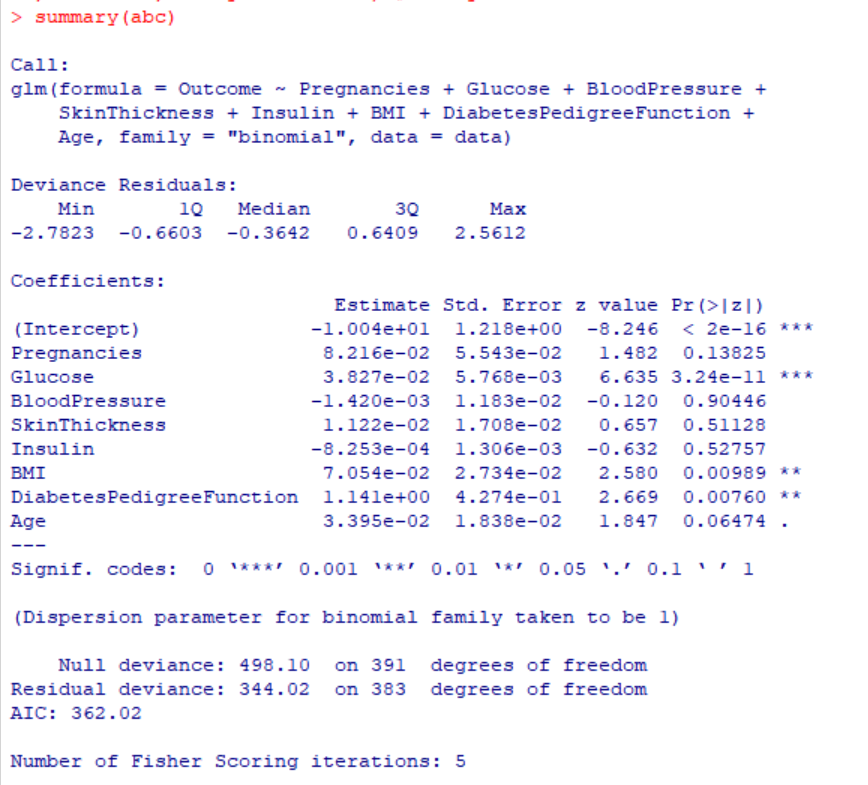
임신 횟수가 많을수록 당뇨병일 확률이 높습니다.

#글루코스



글루코스에 따른 분포가 확실하게 차이가 나 높으면 당뇨병일 확률이 큽니다.

#이분형 로지스틱 회귀분석



Residuals -> 잔차(회귀식에 의해 추정된 값과 실제값의 차이)

Coefficients -> 추정된 회귀식의 계수

Estimate /Std.Error/z value/Pr(>lzl)

추정된 계수/표준오차/z값/p-value

Intercept -> 각 변수의 회귀 직선에 대한 y절편의 합

Null deviance -> 기본 회귀분석에서 전체제곱합

Residual deviance -> 기본 회귀분석의 잔차제곱합

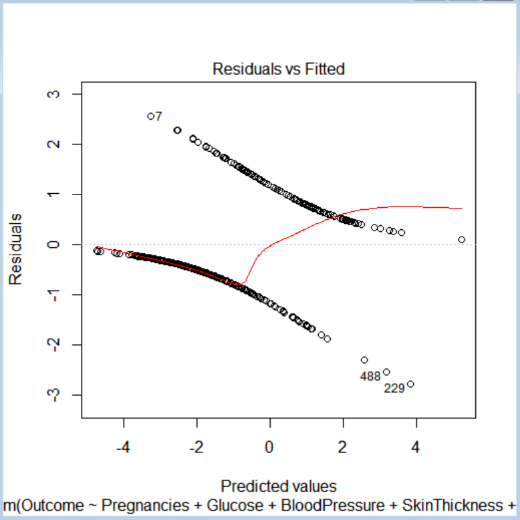
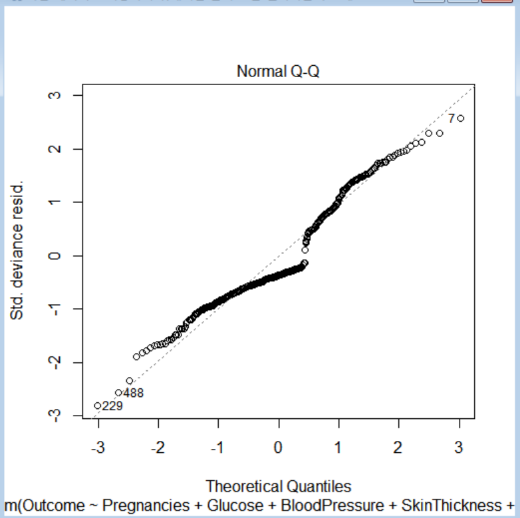
AIC -> 모형의 설명력을 나타내는 척도, 결정계수나 조정된 결정계수 역할

Number of Fisher Scoring iterations -> 최종 모형으로 수렴할 때까지 반복횟수

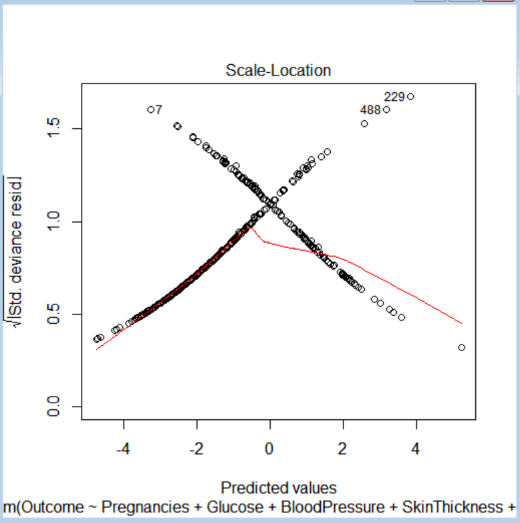
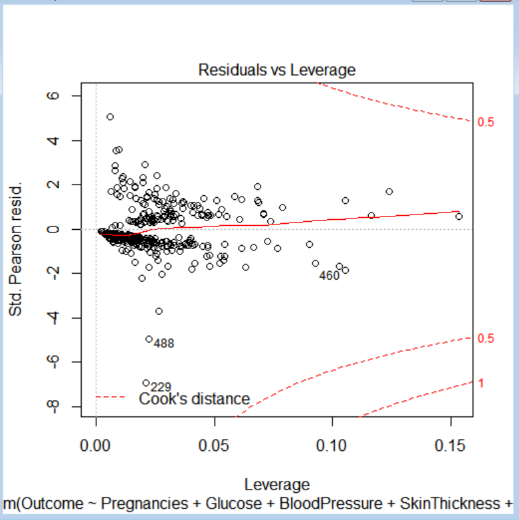
\*유의수준을 0.05로 잡아서 p-value와 비교

:Glucose, BMI, DiabetesPedigreeFunction 이 당뇨병과 관련성있다고 말할 수 있습니다.

#회귀직선....?

X축에는 선형 회귀로 예측된 Y값, Y축에는 잔차 정규분포를 따르는지 검토

X축에는 선형 회귀로 예측된 Y값, Y축에는 표준화 잔차 X축에 레버리지, Y축에 표준화 잔차